

The temporal window of audio-tactile integration in speech perception

Bryan Gick, Yoko Ikegami, and Donald Derrick

*Department of Linguistics, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada
gick@interchange.ubc.ca, yoko@ikegami.jp, dderrick@interchange.ubc.ca*

Abstract: Asynchronous cross-modal information is integrated asymmetrically in audio-visual perception. To test whether this asymmetry generalizes across modalities, auditory (aspirated “pa” and unaspirated “ba” stops) and tactile (slight, inaudible, cutaneous air puffs) signals were presented synchronously and asynchronously. Results were similar to previous AV studies: the temporal window of integration for the enhancement effect (but not the interference effect) was asymmetrical, allowing up to 200 ms of asynchrony when the puff followed the audio signal, but only up to 50 ms when the puff preceded the audio signal. These findings suggest that perceivers accommodate differences in physical transmission speed of different multimodal signals.

© 2010 Acoustical Society of America

PACS numbers: 43.71.Rt, 43.71.Es [AL]

Date Received: May 17, 2010 Date Accepted: September 29, 2010

1. Introduction

It is well known that asynchronously presented auditory and visual information is integrated asymmetrically in speech perception (Dixon and Spitz, 1980; Smeele *et al.*, 1992; Summerfield, 1992). For example, Munhall *et al.* (1996) found that audio-visual integration of speech occurred even when the audio signal lagged the video signal by 240 ms; however, when the audio signal preceded the video signal, perceivers only integrated with 60 ms or less of asynchrony. They conjecture that this asymmetrical effect window may be attributable to perceivers’ learned awareness of physical properties of the natural world (in this case, of the differing atmospheric speeds of sound and light): “This trend is not surprising since the relative speeds of sound and light would produce many natural occurrences of auditory events lagging their visual counterparts in the natural world” (Munhall *et al.*, 1996, p. 354), suggesting that human perceptual systems include a learned or innate awareness of the laws of physics. This explanation for the asymmetry in audio-visual perception has not, however, been substantiated via comparison with other pairs of perceptual modalities. Replication using the tactile modality would provide a test case for this question.

Fowler and Dekle (1991) and Gick *et al.* (2008) found that untrained perceivers integrate tactile and auditory modalities through direct manual contact with speakers’ faces. However, even if realistic and precisely timed synthetic facial (e.g., robotic) stimuli could be constructed, this methodology would still fail to provide a natural signal transmission delay comparable to that of light or sound, as direct manual information and proximate acoustical information are always received approximately simultaneously.

In a recent study, Gick and Derrick (2009) used small puffs of air to influence auditory speech perception. Participants who received puffs of air on their necks or hands while simultaneously hearing aspirated or unaspirated English plosives (i.e., “pa” or “ba”) were more likely to perceive both sounds as aspirated (that is, “pa”), suggesting that listeners integrate this tactile and auditory speech information in much the same way as they do synchronous visual and auditory information. Further, we know that the air speed of the turbulent flow released in speech aspiration is considerably slower than that of sound in air, with flow velocity dropping

off log-linearly after expulsion from the mouth (Derrick *et al.*, 2009). Thus, this combination of stimuli provides both the mechanism for synthesis of stimuli and the natural temporal delay needed for the present study.

The present experiment follows the air-puff methodology, coupling an acoustic speech signal with small bursts of air on the skin, but delivered with a range of positive or negative temporal offsets. If the physics-based hypothesis (i.e., the explanation based on perceivers' awareness of the relative physical transmission times of different signals) is correct, then the direction of asymmetry in the perceptual integration window should parallel the temporal difference between the relative velocities of sound and air flow. Specifically, we predict that perceivers will continue to integrate the two signals despite longer temporal offsets when air-puffs (the slower signal) follow acoustics (the faster signal), while perceivers will cease to integrate when air-puffs precede the acoustic signal by a substantial margin.

2. Method

Thirteen adult perceivers participated in the study. All were right-handed, native speakers of English with no prior phonetics training, and no history of speech or hearing problems.

Acoustic stimuli consisted of recordings of 440 tokens of *pa* and *ba* naturally produced by a single female English speaker and presented in random order. Acoustic stimuli were output through the right channel of a Mac G4 sound card, mixed through a PreSonus mixing board with white noise (at a level such that subjects' baseline correct identification of *pa/ba* was at approximately 75%) and played to participants in stereo through Direct Sound Extreme Isolation headphones.

Tactile stimuli consisted of gentle bursts of air imparted via a vinyl tube at 7 cm from the neck, to the right of the suprasternal notch. Bursts were released from an air compressor at ~5 psi using a Teknocrat 12-V dc 2-way solenoid valve with a 0.032-in. orifice. The switch operating the solenoid valve was activated by a voltage initiated by an acoustic square wave output through the left channel a Mac G4 sound card amplified to 5 V using a Frequency Devices voltage amplifier. Square waves were 60 ms long (the average duration of aspiration for the natural "pa" tokens used in the experiment), and offset leftward by 30 ms to correct for a total system latency of 30 ms (see Fig. 1). By comparison, for English word-onset voiceless (aspirated) stops, average aspiration duration is around 54–80 ms, and average air pressure is up to 7 cm H₂O (Lisker and Abramson, 1964), with pressure received at the skin decreasing logarithmically with distance from the source (Derrick *et al.*, 2009).

Twenty-four experimental conditions were tested, with spoken tokens "pa" or "ba" paired with air bursts at temporal offsets as follows: No Burst, 0 ms (Simultaneous), ± 50 , ± 100 , ± 200 , ± 300 , and ± 500 ms. The positive durations correspond to aspiration at about 17 cm (50 ms), 24 cm (100 ms), 31 cm (200 ms), 36 cm (300 ms), and 42 cm (500 ms) (Derrick *et al.*, 2009). Aspiration-related air flow becomes largely dissipated between 30 and 40 cm away, so the 500 ms tokens are too temporally distant for aspiration and therefore represent distractors. Participants heard 20 items for each experimental condition and 10 items each for the two distractor conditions, presented in random order, with one item occurring every 3 s.

Participants were seated in a sound booth and were read a script describing this experiment as testing their ability to identify different spoken syllables under conditions similar to those experienced by an airplane pilot. No specific mention was made of the air tube (indeed, some subjects reported not being aware of the air burst at all during the experiment). Participants were briefly instructed in making forced-choice responses using a button box (with L/R responses balanced across participants), then blindfolded. Headphones were then placed on the participant, and the air tube put in place aiming at the subject's neck, to the right of the suprasternal notch.

3. Results

Figure 2 compares the mean percentage of correctly identified "pa" and "ba" syllables in No puff vs. Synchronous puff conditions. When "pa" and "ba" were coupled with a synchronous burst (i.e., in Simultaneous conditions), paired t-tests (by subject) showed significant enhance-

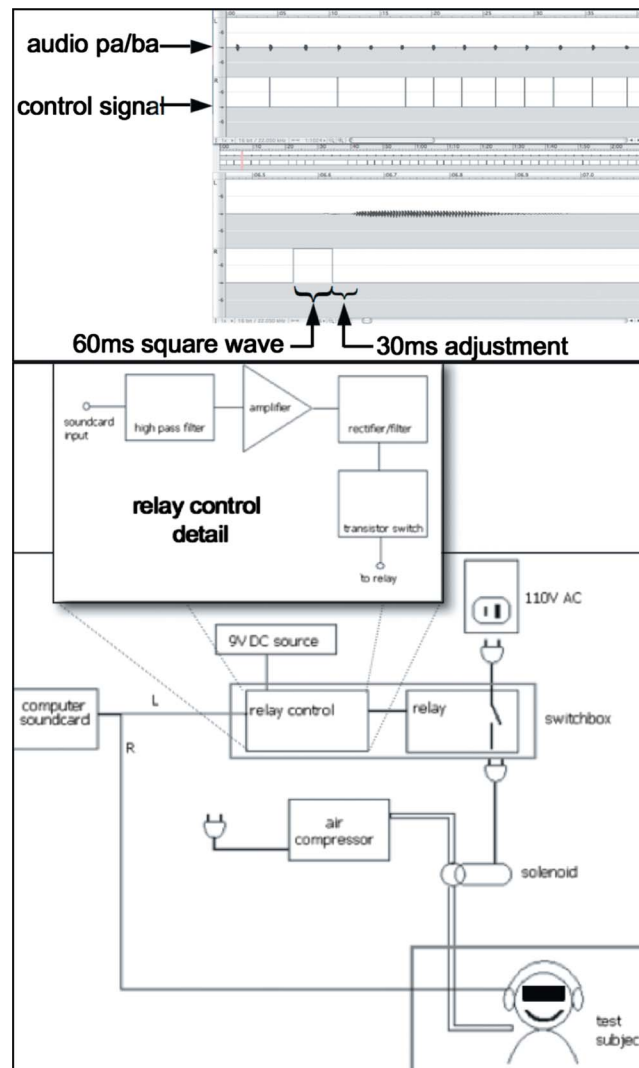


Fig. 1. (Color online) Example of acoustic control and *pa/ba* signals (top) and flowchart of stimulus presentation system (bottom).

ment to identification of “pa” responses [$t(12)=-2.2592, p=0.04$], and paired t-tests showed significant interference with identification of “ba” responses [$t(12)=2.63, p=0.02$], when compared with No Burst baseline conditions. Figure 3 shows the mean percentage of correctly identified “pa” and “ba” syllables across subjects, plotted by temporal offset condition. For both “pa” and “ba,” the effect at -50 ms was not significantly different from simultaneous, {“ba” t-test [$t(12)=0.74, p=0.48$], “pa” t-test [$t(12) \sim 0, p \sim 1$]}, but the effect at -100 ms was different from simultaneous {“ba” t-test [$t(12)=2.33, p=0.04$], “pa” t-test [$t(12)=-2.41, p=0.03$]}. In the positive offset direction, while the effect for “ba” mirrored that of the negative offset direction, continuing to show integration at $+50$ ms [$t(12)=-0.09, p=0.93$] but not at $+100$ ms [$t(12)=-2.96, p=0.01$], integration for “pa” persisted at delays of $+50$ ms [$t(12)=0.14, p=0.89$], $+100$ ms [$t(12)=0.29, p=0.77$], and $+200$ ms [$t(12)=1.29, p=0.22$].

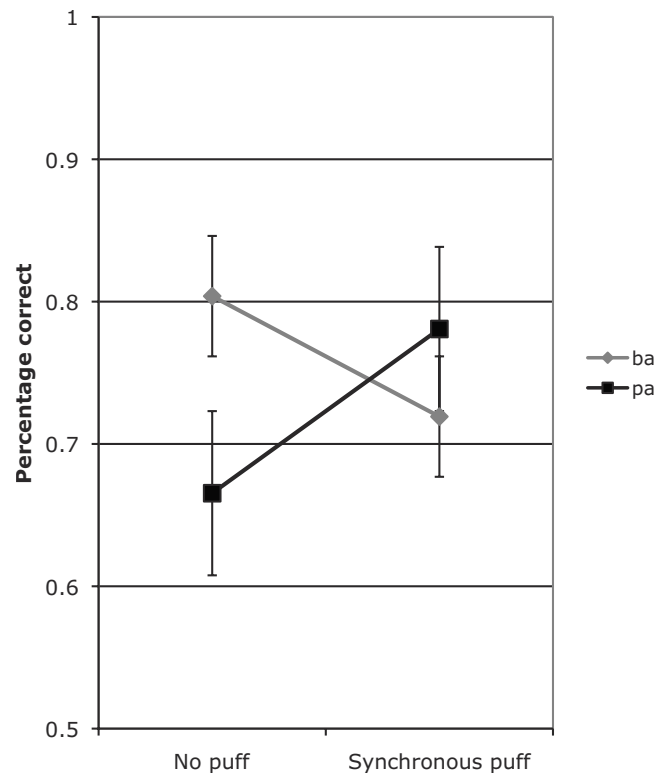


Fig. 2. Mean percentage of correctly identified “pa” (black line) and “ba” (gray line) syllables comparing No puff vs. Synchronous puff.

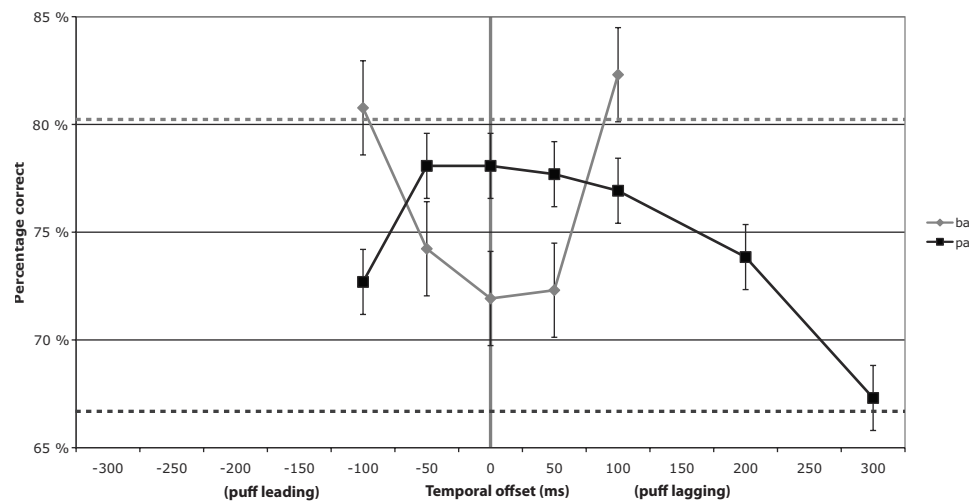


Fig. 3. Temporal window of integration showing mean percentage of correctly identified “pa” (black line) and “ba” (gray line) syllables. The vertical line indicates the zero-offset point, intersecting the Synchronous puff condition. Horizontal dotted lines show baseline percentage accuracy without puff, with black indicating “pa,” and gray indicating “ba,” i.e., contiguous significant data points centering on a zero offset, ending with the first non-significant result.

4. Discussion

In this experiment, a burst occurring immediately prior to vowel onset (i.e., synchronous with normal aspiration for “pa”) significantly enhanced perception of “pa” and significantly interfered with perception of “ba,” replicating Gick and Derrick (2009).

Asynchronous results differed for interference (“ba”) vs. enhancement (“pa”). Results for “ba” showed a narrow, symmetrical window of integration, of ± 50 ms. Results for “pa,” however, showed an asymmetrical effect window similar to that observed in previous studies of audio-visual integration: for “pa,” integration continued to occur when the air burst followed the audio signal by up to 200 ms, but only by 50 ms when the air burst preceded the audio signal.

An unexpected additional finding is that the maximum delay at which integration occurred under our laboratory conditions (200 ms) corresponds with the maximum time window during which an actual speech-related air puff would be perceivable due to flow dissipation under typical atmospheric conditions (between 200 and 300 ms; Derrick *et al.*, 2009).

5. Conclusion

The direction of the perceptual asymmetry observed in cross-modal enhancement in this study parallels the temporal difference between the speeds of sound and air flow, supporting the hypothesis that our perceptual systems incorporate physical laws. Future work may determine why the interference effect does not exhibit the expected asymmetrical window, and whether perceivers’ apparent understanding of physical properties of the world is learned or innate.

Acknowledgments

This project benefited greatly from technical contributions of Gordon Ramsay, assistance from Leonardo Oliveira, and discussions with Douglas Whalen, and was funded by an NSERC Discovery Grant to Bryan Gick and NIH Grant No. DC-02717 to Haskins Laboratories.

References and links

- Derrick, D., Anderson, P., Gick, B., and Green, S. (2009). “Characteristics of air puffs produced in English ‘pa’: Experiments and simulations,” *J. Acoust. Soc. Am.* **125**, 2272–2281.
- Dixon, N., and Spitz, L. (1980). “The detection of audiovisual desynchrony,” *Perception* **9**, 719–721.
- Fowler, C. A., and Dekle, D. J. (1991). “Listening with eye and hand: Cross-modal contributions to speech perception,” *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 816–828.
- Gick, B., and Derrick, D. (2009). “Aero-tactile integration in speech perception,” *Nature (London)* **462**, 502–504.
- Gick, B., Jóhannsdóttir, K., Gibrael, D., and Muehlbauer, J. (2008). “Tactile enhancement of auditory and visual speech perception in untrained perceivers,” *J. Acoust. Soc. Am.* **123**, EL72–EL76.
- Lisker, L., and Abramson, A. S. (1964). “A cross-language study of voicing in initial stops: acoustical measurements,” *Word* **20**, 384–423.
- Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (1996). “Temporal constraints on the McGurk effect,” *Percept. Psychophys.* **58**, 351–362.
- Smeele, P. M. T., Sittig, A. C., and Van Heuven, V. J. (1992). “Intelligibility of audio-visually desynchronized speech: Asymmetrical effect of phoneme position,” in *Proceedings of the International Conference on Spoken Language Processing*, pp. 65–68.
- Summerfield, Q. (1992). “Lipreading and audio-visual speech perception,” *Philos. Trans. R. Soc. London, Ser. B* **335**, 71–78.